산업공학특론I 3주차_확률변수와 확률분포 실습

임문원 (moonmunwon@psm.hanyang.ac.kr (mailto:moonmunwon@psm.hanyang.ac.kr))

2024-03-20

[시뮬레이션]

1. 중심극한정리

```
x <- list()
for (i in 1:100){
  x[[i]] <- rnorm(10000, mean=20, sd=10)
}
meanx <- Reduce('+',x)/length(x)
cat(mean(meanx), sd(meanx))
```

```
## 19.99155 1.003963
```

2. 평균, 분산

```
meanvar <- function(dist, n=10000, param){
  if (dist=='binom'){
    rand <- rbinom(n, param[1], param[2])
    m <- param[1]*param[2]; v <- param[1]*param[2]*(1-param[2])
  } else if (dist=='expo'){
    rand <- rexp(n, param[1])
    m <- 1/param[1]; v <- 1/param[1]^2
  } else if (dist=='chisq'){
    rand <- rchisq(n, param[1])
    m <- param[1]; v <- 2*param[1]
  }
  print(paste('시뮬레이션:',mean(rand),var(rand)))
  print(paste('산출공식  :',m,v))
}

meanvar(dist='binom',param=c(10,0.1))
```

```
## [1] "시뮬레이션: 1.0087 0.911915501550155"
## [1] "산출공식  : 1 0.9"
```

```
meanvar(dist='expo', param=0.1)
```

```
## [1] "시뮬레이션: 10.0017471659303 100.358555983513"
## [1] "산출공식  : 10 100"
```

```
meanvar(dist='chisq', param=10)
```

```
## [1] "시뮬레이션: 10.0274607143929 20.2427458897035"
## [1] "산출공식  : 10 20"
```

[데이터 분석]

0. 대상 데이터 확보: 공구 마모 데이터셋

(https://www.kaggle.com/datasets/shivamb/machine-predictive-maintenance-classification
(https://www.kaggle.com/datasets/shivamb/machine-predictive-maintenance-classification))

생산 공정의 예지보전을 목적으로 실제 측정 데이터에 공정의 환경을 합성한 데이터

다중 공정 중 단일한 설비를 대상으로 데이터 수집/생성

10,000개의 데이터 포인트와 10개의 변수로 구성되며, 각 변수에 대한 설명은 다음과 같음

- UID: 1부터 10,000까지의 범위를 가지는 고유 식별자
- ProductID: 제품 등급에 따라 낮음(L), 중간(M), 높음(H)으로 분류하였으며, 각각 전체의 50%, 30%, 20%를 차지
- Air temperature [K]: 표준 편차가 2K인 300K 주변으로 정규화된 랜덤 워크 과정을 사용하여 인위적으로 생성
- Process temperature [K]: 표준 편차가 1K인 랜덤 워크 과정을 사용하여 생성되고, 공기 온도에 10K를 더한 값에 정규화
- Rotational speed [rpm]: 2860W를 중심으로 계산 및 생성되었으며, 정규 분포를 따르는 잡음이 인가되어 있음
- Torque [Nm]: 40Nm 주변에서 정규분포를 따르며, 표준 편차는 10Nm이고, 음수 값은 없음
- Tool wear [min]: 공구의 마모 시간으로, Target 변수와 연동하여 고장 유무에 따른 시간을 검토할 수 있음
- Target: 고장 유무
- Failure type: 고장의 유형

1. 데이터 탐색 (EDA) 및 전처리

```
# 데이터 로드 및 조회
dat <- read.csv('산업공학특론I_3주차_실습 데이터.csv')
head(dat)
```

```
##   UDI Product.ID Type Air.temperature..K. Process.temperature..K.
## 1   1     M14860    M               298.1                   308.6
## 2   2     L47181    L               298.2                   308.7
## 3   3     L47182    L               298.1                   308.5
## 4   4     L47183    L               298.2                   308.6
## 5   5     L47184    L               298.2                   308.7
## 6   6     M14865    M               298.1                   308.6
##   Rotational.speed..rpm. Torque..Nm. Tool.wear..min. Target Failure.Type
## 1                   1551        42.8               0      0   No Failure
## 2                   1408        46.3               3      0   No Failure
## 3                   1498        49.4               5      0   No Failure
## 4                   1433        39.5               7      0   No Failure
## 5                   1408        40.0               9      0   No Failure
## 6                   1425        41.9              11      0   No Failure
```

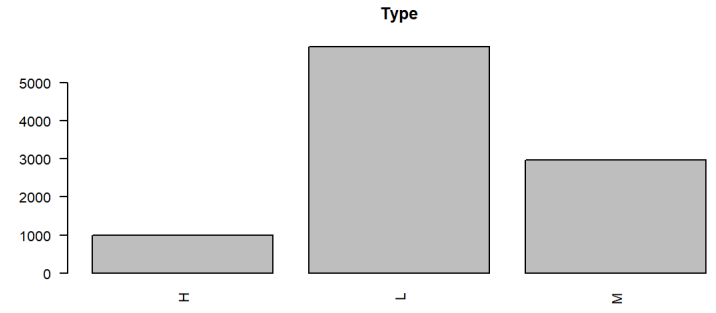
```
# 필요하지 않은 변수 삭제 / 데이터 전처리
dat <- dat[,-(1:2)]

# 필요하지 않은 변수 내 값들 제거
# (Failure.Type 변수 내에 존재하는 Failure, Failures 값은 고장 유형 구분 시 큰 도움이 되지 않는 값이므로 삭제)
dat$Failure.Type <- gsub(' Failure| Failures','', dat$Failure.Type)
dat$Failure.Type <- as.factor(dat$Failure.Type)

# 스페이스로 이루어진 변수에 '.'이 포함되어 가독성을 저해하므로 제외
colnames(dat) <- gsub('[.]','',colnames(dat))
summary(dat)
```

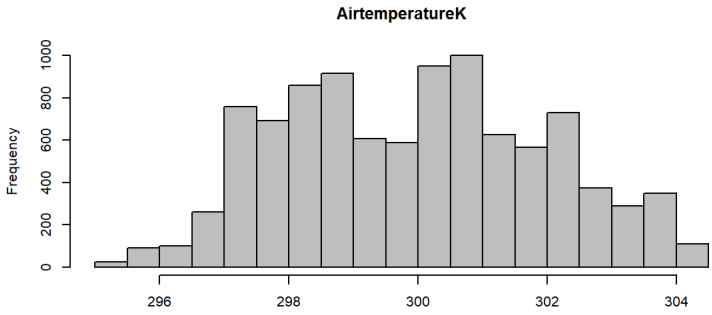
```
##      Type           AirtemperatureK ProcesstemperatureK Rotationalspeedrpm
##  Length:10000       Min.   :295.3   Min.   :305.7       Min.   :1168
##  Class :character   1st Qu.:298.3   1st Qu.:308.8       1st Qu.:1423
##  Mode  :character   Median :300.1   Median :310.1       Median :1503
##                     Mean   :300.0   Mean   :310.0       Mean   :1539
##                     3rd Qu.:301.5   3rd Qu.:311.1       3rd Qu.:1612
##                     Max.   :304.5   Max.   :313.8       Max.   :2886
##     TorqueNm      Toolwearmin      Target                FailureType
##  Min.   : 3.80   Min.   :  0   Min.   :0.0000   Heat Dissipation: 112
##  1st Qu.:33.20   1st Qu.: 53   1st Qu.:0.0000   No              :9652
##  Median :40.10   Median :108   Median :0.0000   Overstrain      :  78
##  Mean   :39.99   Mean   :108   Mean   :0.0339   Power           :  95
##  3rd Qu.:46.80   3rd Qu.:162   3rd Qu.:0.0000   Random          :  18
##  Max.   :76.60   Max.   :253   Max.   :1.0000   Tool Wear       :  45
```

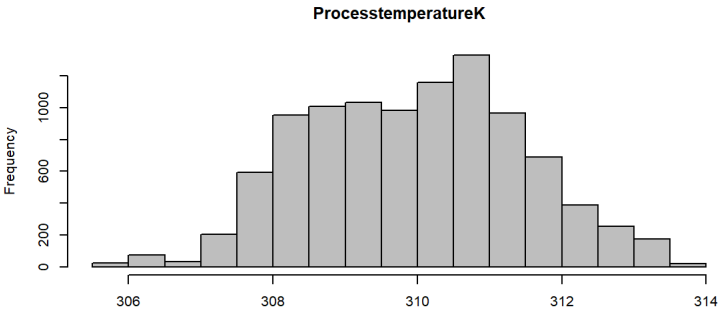
```
# 마모 시간이 0 이하인 데이터 제거
# (강의시간에 언급했던 감마분포의 정의역 (t>0)을 고려하기 위함)
dat <- dat[dat$Toolwearmin>0,]

# 고장여부에 대한 구분자를 이해하기 쉽게 변경
dat$Target <- ifelse(dat$Target==1,'Fail','Normal')

# 문자형 변수들을 범주형으로 변경
dat$Target <- as.factor(dat$Target)
dat$Type <- as.factor(dat$Type)
dat$FailureType <- as.factor(dat$FailureType)

summary(dat)
```

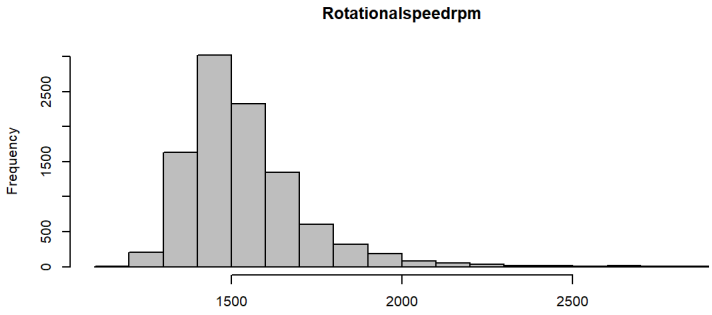
```
##  Type     AirtemperatureK ProcesstemperatureK Rotationalspeedrpm
##  H: 986   Min.   :295.3   Min.   :305.7       Min.   :1168
##  L:5931   1st Qu.:298.3   1st Qu.:308.8       1st Qu.:1423
##  M:2963   Median :300.1   Median :310.1       Median :1503
##           Mean   :300.0   Mean   :310.0       Mean   :1539
##           3rd Qu.:301.5   3rd Qu.:311.1       3rd Qu.:1612
##           Max.   :304.5   Max.   :313.8       Max.   :2886
##     TorqueNm      Toolwearmin       Target              FailureType
##  Min.   : 3.80   Min.   :  2.0   Fail  : 336   Heat Dissipation: 112
##  1st Qu.:33.20   1st Qu.: 55.0   Normal:9544   No              :9535
##  Median :40.10   Median :109.0                 Overstrain      :  78
##  Mean   :39.98   Mean   :109.3                 Power           :  92
##  3rd Qu.:46.80   3rd Qu.:163.0                 Random          :  18
##  Max.   :76.60   Max.   :253.0                 Tool Wear       :  45
```

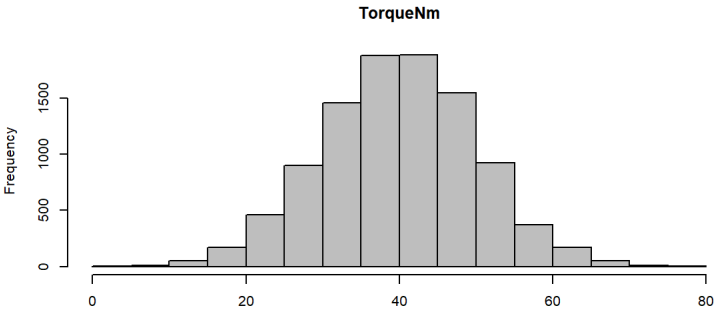
```
# 데이터 탐색을 위한 시각화 수행
# 이 때, 범주형은 막대그래프, 연속형은 히스토그램으로 분포 시각화하도록 함수 작성
visualize <- function(x,main){
  if (is.factor(x)){ barplot(table(x), col='grey', main=main, las=2)
  } else { hist(x, col='grey', main=main, xlab='') }
}

# 4행 2열 배치로 시각화 진행
par(mfrow=c(4,2))
for (i in 1:ncol(dat)){ visualize(x=dat[,i], main=colnames(dat)[i]) }
```

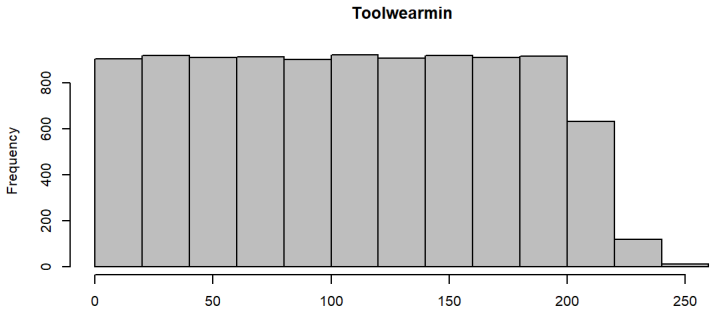


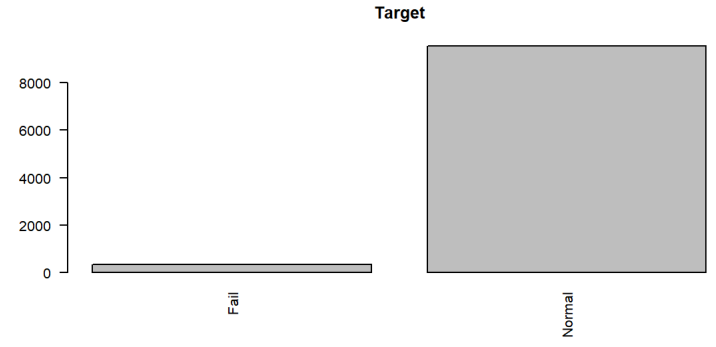


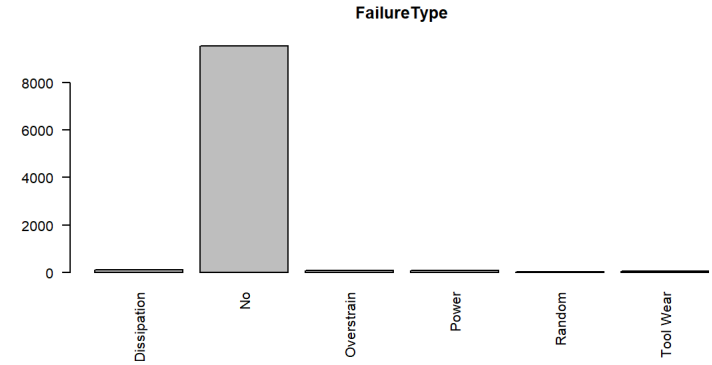












2. 데이터 그룹화 및 기술통계량 검토

```
# 타입에 따라 그룹화된 데이터 리스트 생성 및 구분된 그룹별 기술통계량 검토
# lapply 함수를 사용하면 리스트 단위의 개별 summary를 진행할 수 있음

# 제품 품질 (Type)에 따른 마모량 (Wear) 분포 확인
dat_product <- split(dat$Toolwearmin, dat$Type)
lapply(dat_product, summary)
```

```
## $H
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   55.25  109.00  109.27  161.00  246.00
##
## $L
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0    55.0   110.0   109.6   164.0   251.0
##
## $M
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0    55.0   107.0   108.5   163.0   253.0
```

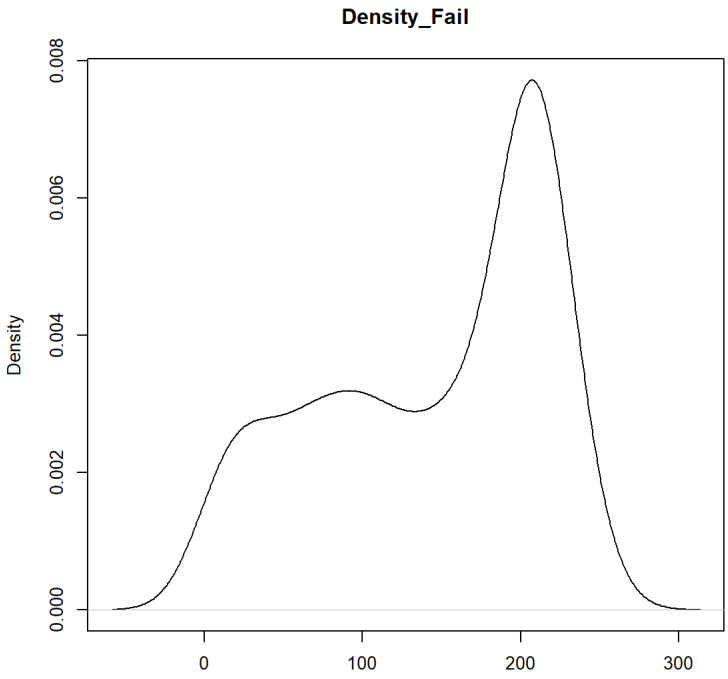
```
# 고장 유형 (Failure Type)에 따른 마모량 (Wear) 분포 확인
dat_failure <- split(dat$Toolwearmin, dat$FailureType)
lapply(dat_failure,summary)
```

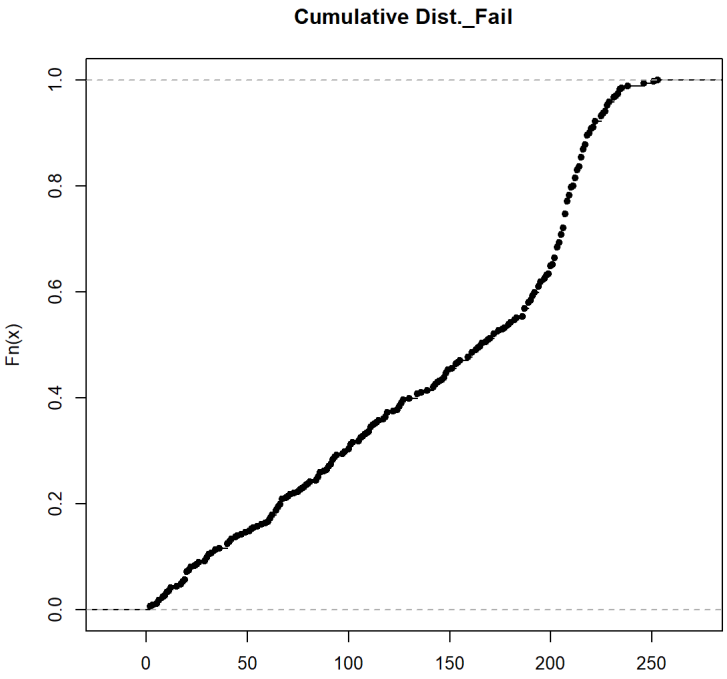
```
## $`Heat Dissipation`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0    54.5   106.0   107.3   161.5   229.0
##
## $No
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       2      54     108     108     161     246
##
## $Overstrain
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   177.0   200.0   207.0   208.2   216.0   251.0
##
## $Power
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0    60.0   101.0   105.2   151.5   234.0
##
## $Random
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   61.75  142.00  119.89  171.50  215.00
##
## $`Tool Wear`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   198.0   207.0   215.0   216.6   225.0   253.0
```

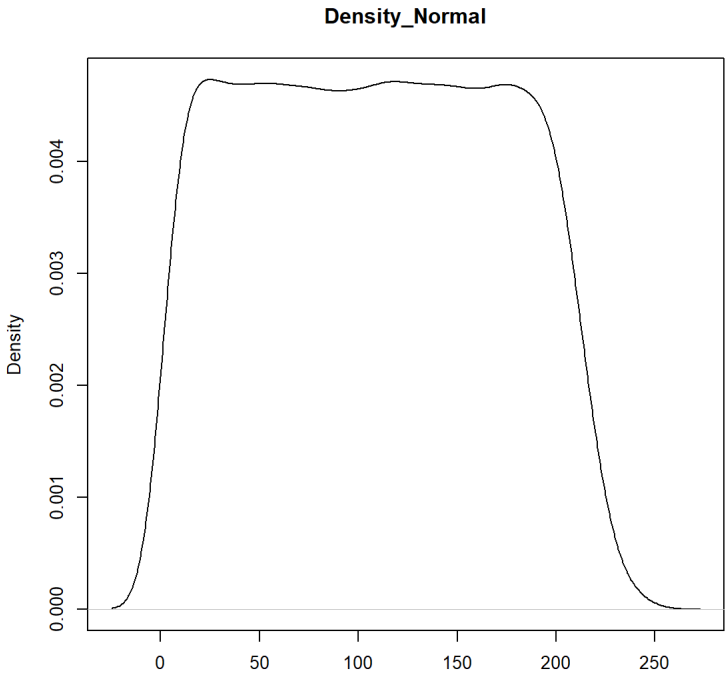
```
# 고장 여부 (Target)에 따른 마모량 (Wear) 분포 확인
dat_target <- split(dat$Toolwearmin, dat$Target)
lapply(dat_target,summary)
```

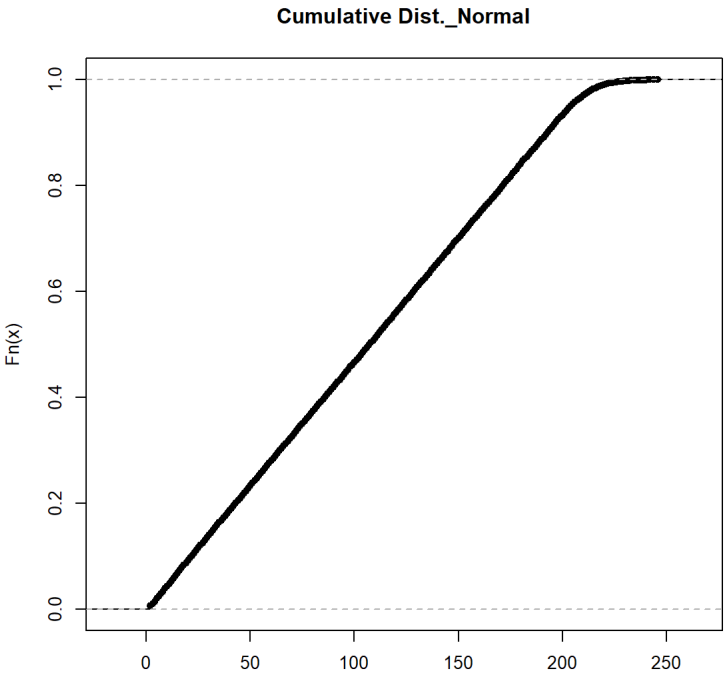
```
## $Fail
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   85.75  166.00  145.07  208.00  253.00
##
## $Normal
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       2      54     108     108     161     246
```

```
# 관측 변수 기반 유형별 경험적 밀도/누적분포 시각화 함수 정의
dens <- function(x){
  par(mfrow=c(length(x),2))
  for (i in 1:length(x)){
    plot(density(x[[i]]), main= paste('Density_',names(x)[i],sep=''),xlab='')
    plot(ecdf(x[[i]]), main=paste('Cumulative Dist._',names(x)[i],sep=''), xlab='')
  }
}

# 우리가 살펴보고자 하는 고장 여부 (Target)에 따른 마모량 (Wear) 분포에 대한 밀도/누적분포 도출
dens(dat_target)
```









3. 확률분포 적합

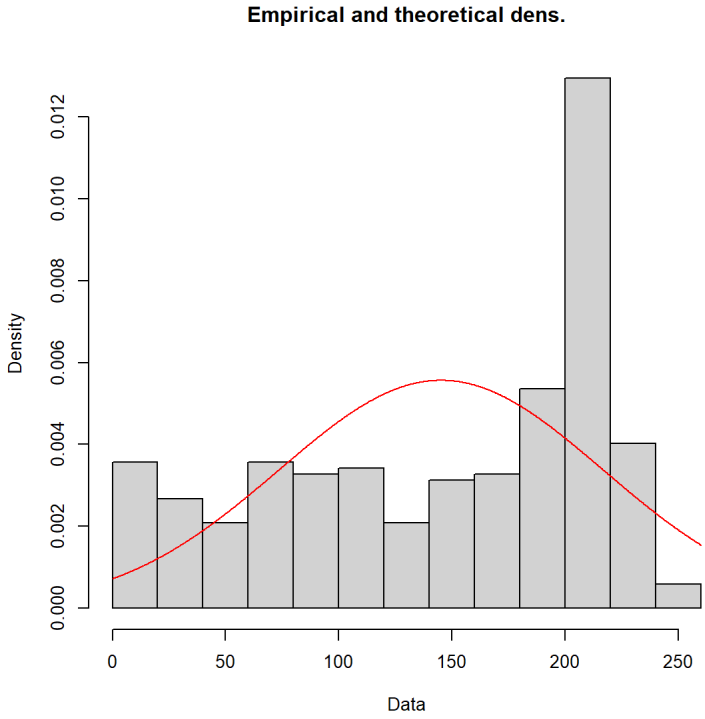
```
# 확률분포 관련 라이브러리 로드
library(fitdistrplus)
```

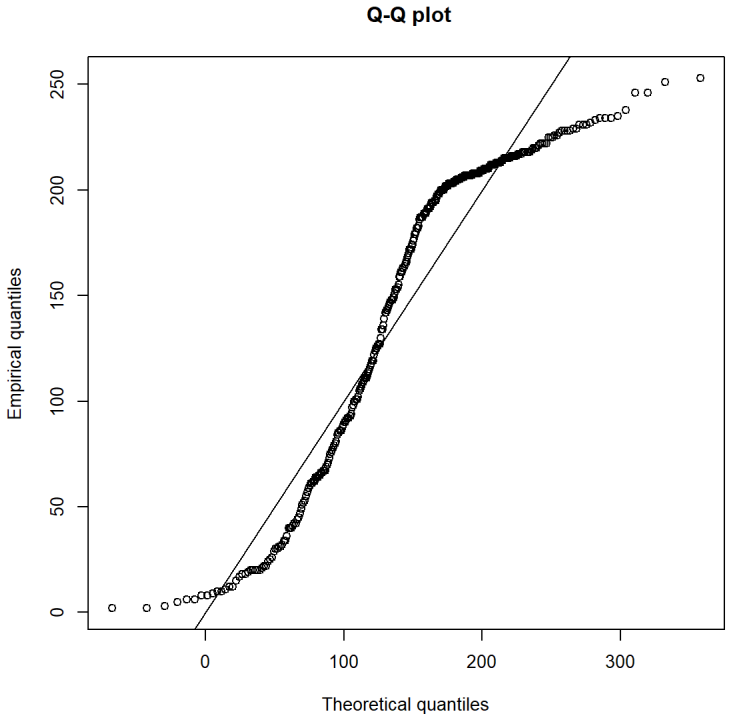
```
## 필요한 패키지를 로딩중입니다: MASS
```

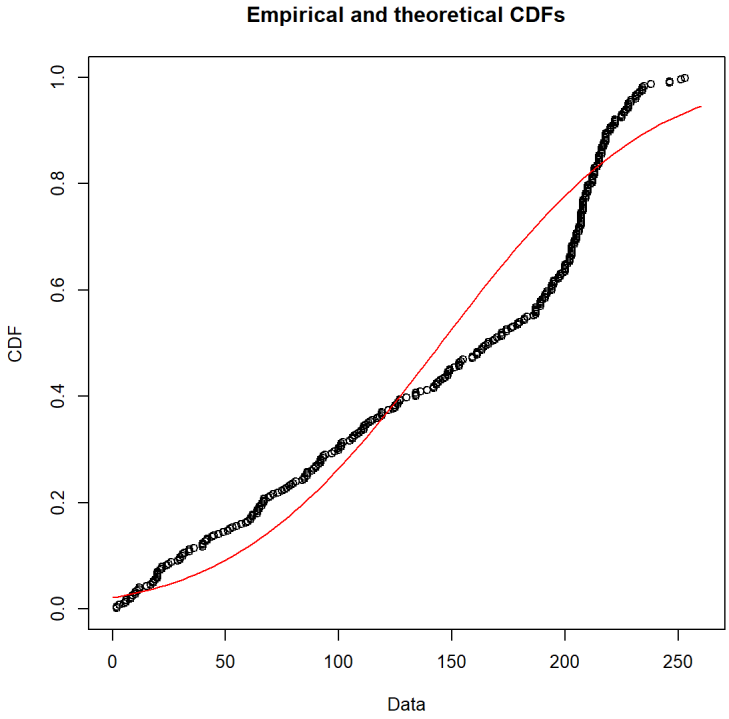
```
## 필요한 패키지를 로딩중입니다: survival
```

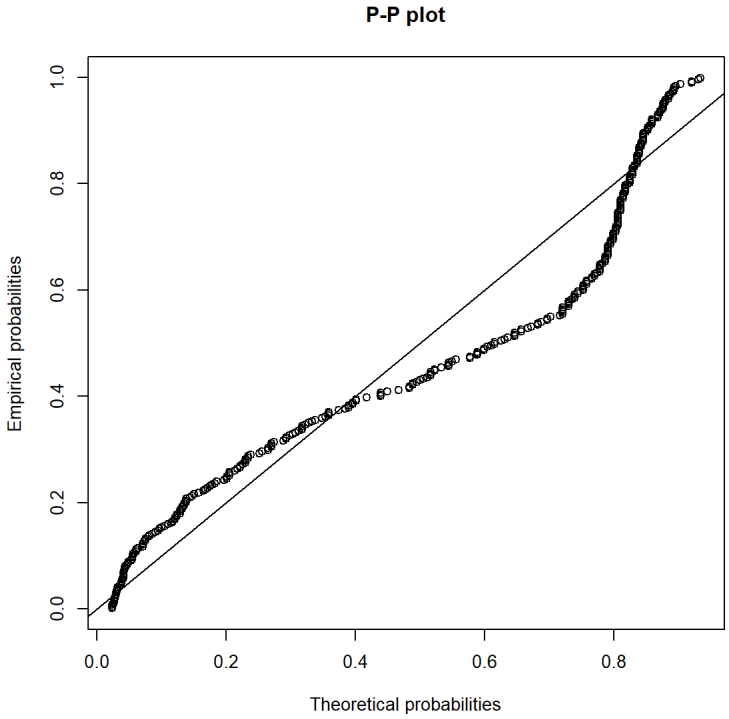
```
# 확률분포 추정 함수 정의
fit <- function(x){

  # 후보 확률분포 정의
  fitlist <- c('exp','gamma','norm')

  # 모든 그룹에 대한 확률변수 추정 진행
  for (i in 1:length(x)){
    print( '#########################################################' )
    print( paste(names(x)[i],'그룹에 대한 추정을 시작합니다.') )

    # 추정 파라미터와 로그우도값을 추정한 결과를 정리할 틀 생성
    result <- matrix(nrow=3, ncol=length(fitlist))
    colnames(result) <- fitlist
    row.names(result) <- c('par1','par2','loglike')

    # 지정한 후보 확률분포를 각각 추정하여 결과를 틀에 입력
    for (j in 1:length(fitlist)){
      fit_temp <- fitdist(x[[i]], fitlist[j])

      est <- fit_temp$estimate
      if (length(est)==1){ est <- c(est,NA)}

      ll <- fit_temp$loglik
      result[,j] <- c(est, ll)
    }
    print(result) # 결과 출력

    # 로그우도값이 가장 큰 확률분포를 적합 분포로 선정하여 출력
    # (우리는 최대우도추정법을 기반으로 모수를 추정하는 관계로, 로그우도값이 큰 값이 나오는 분포가 데이터를 잘 설명한
다고 판단
    # 최우추정법에 대한 내용은 추정 시간에 다룰 예정이니 참고)
    best <- colnames(result)[which.max(result[3,])]
    print(paste(names(x)[i],'그룹의 최적 분포는 ', best,'입니다.'))
    plot(fitdist(x[[i]], best))
  }
}

fit(dat_target)
```

```
## [1] "#########################################################"
## [1] "Fail 그룹에 대한 추정을 시작합니다."
##                   exp         gamma        norm
## par1     6.893439e-03     2.1646650   145.06548
## par2               NA     0.0149202    71.68878
## loglike -2.008334e+03 -1961.4373949 -1912.26766
## [1] "Fail 그룹의 최적 분포는  norm 입니다."
```

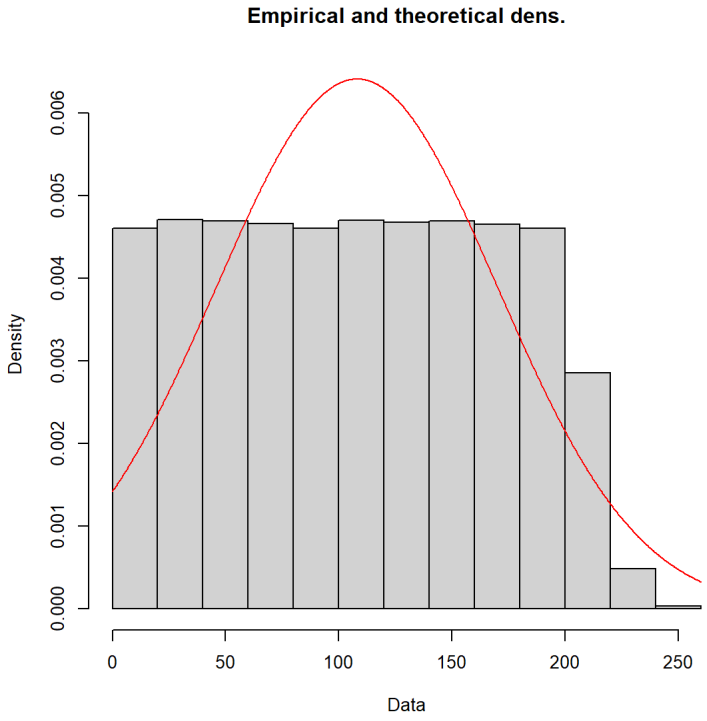


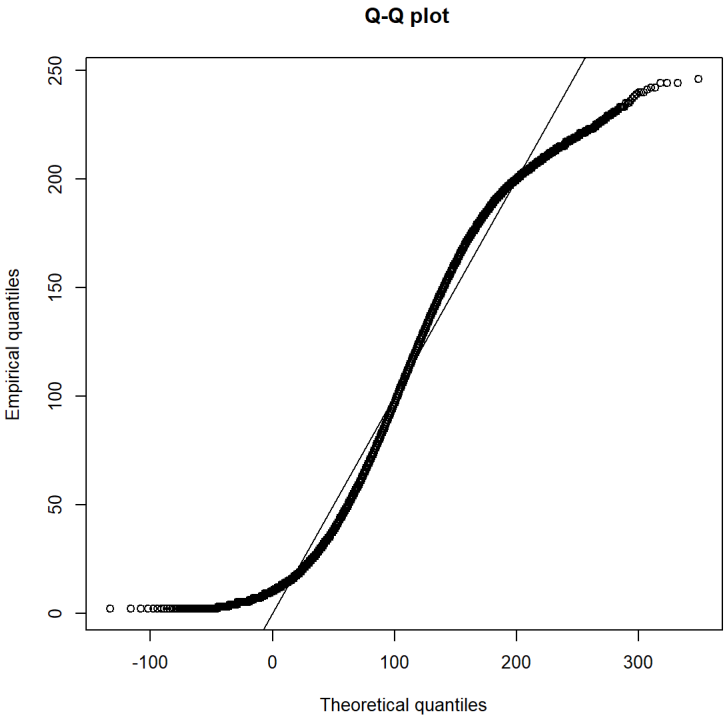


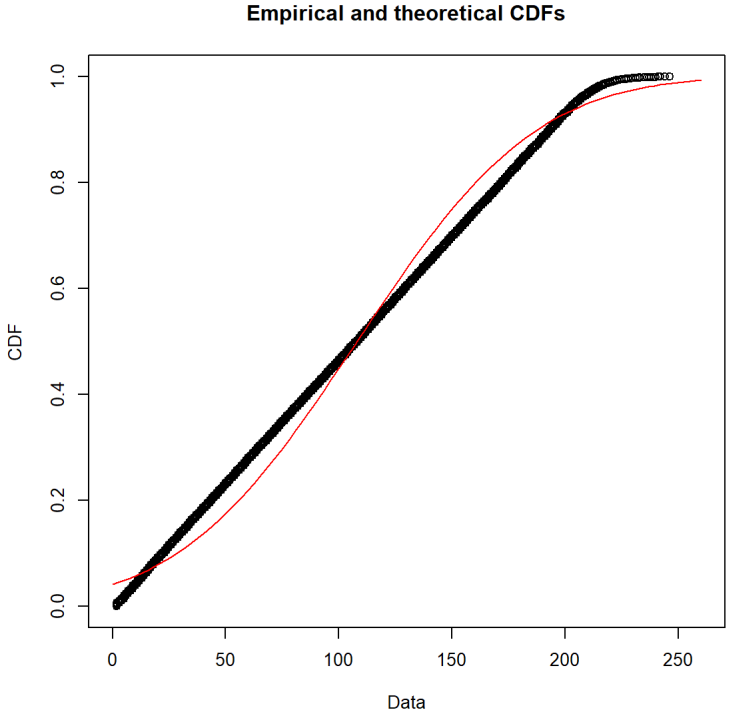


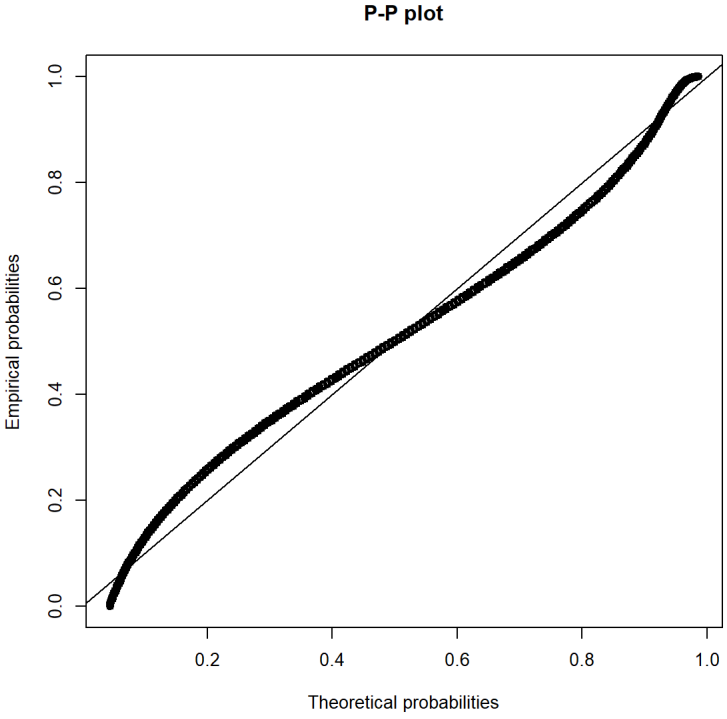



```
## [1] "#########################################################"
## [1] "Normal 그룹에 대한 추정을 시작합니다."
##                   exp         gamma         norm
## par1     9.259116e-03  1.899991e+00    108.00168
## par2               NA  1.759240e-02     62.20183
## loglike -5.423041e+04 -5.326226e+04 -52962.73856
## [1] "Normal 그룹의 최적 분포는  norm 입니다."
```


Empirical and theoretical dens.
Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0
50
100
150
200
250
Data


Q-Q plot
Empirical quantiles
0
50
100
150
200
250
-100
0
100
200
300
Theoretical quantiles


Empirical and theoretical CDFs
CDF
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
Data


P-P plot
Empirical probabilities
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8
1.0
Theoretical probabilities